1-1-1976

Bayes Estimation on Parameters of the Single-Class Classifier

G. C. Lin

T. C. Minter

Reprinted from

Symposium on Machine Processing of Remotely Sensed Data

June 29 - July 1, 1976

The Laboratory for Applications of
Remote Sensing

Purdue University
West Lafayette
Indiana

IEEE Catalog No.
76CH1103-1 MPRSD

BAYES ESTIMATION ON PARAMETERS OF THE SINGLE-CLASS CLASSIFIER*

G. C. Lin and T. C. Minter

Lockheed Electronics Company, Inc./Aerospace Systems Division, Houston, Texas

I. ABSTRACT

Normal procedures used for designing a Bayes classifier to classify wheat as the major crop of interest require not only training samples of wheat but also those of nonwheat. Therefore, ground truth must be available for the class of interest plus all confusion classes. The single-class Bayes classifier classifies data into the class of interest or the class "other" but requires training samples only from the class of interest. This paper will present a procedure for Bayes estimation on the μ_i, Σ_i, q_i (i.e., mean vector, covariance matrix, and a priori probability) of the single-class classifier using labeled samples from the class of interest and unlabeled samples drawn from p(x). The procedure used to derive μ_i, Σ_i, and q_i is to minimize m_L', which is the mean square error of the Bayes decision function of the single-class classifier.

II. INTRODUCTION

The single-class classifier,[1] which needs only training samples of the class of interest, will classify the data into the class of interest or the class "other." The decision rule of the single-class classifier is: Decide $x \in$ wheat if

$$\sum_{j=1}^{m} q_j p(x/j) \geq \frac{1}{2} p(x) \qquad (1)$$

otherwise, $x \in$ "other" where

q_j = the a priori probability of classification as wheat subclass j

$p(x/j)$ = the probability density function of wheat subclass j

$p(x) = \sum_{j=1}^{c} q_j p(x/j)$ = the mixture density function

c = the number of the total subclasses ($c > m$)

This procedure minimizes the need for ground truth; however, the classification performance using the two-class classifier has been shown to be superior to the single-class classifier.[2] When the decision rule of the single-class classifier [equation (1)] was inspected, the estimated mixture density function p(x) was considered a contributing factor to poor classification results. An estimate of p(x) was obtained using clustering.[2] The estimated mixture density function $\hat{p}(x)$ was defined as

$$\hat{p}(x) = \sum_{i=1}^{30} \hat{q}_i \hat{p}(x/i) \qquad (2)$$

where $\hat{p}(x/i)$ is the probability density function of cluster i, which was estimated by clustering the total samples into 30 clusters; $\hat{q}_i$ is N_i/N; N_i is the number of samples which fall into cluster i; and N is the number of total samples. The probability density functions $\hat{p}(x/i)$, $i = 1, 2, \cdots, 30$, are assumed normally distributed with mean vectors μ_i, $i = 1, 2, \cdots, 30$; covariance matrices Σ_i, $i = 1, 2, \cdots, 30$; and the a priori probabilities $\hat{q}_i$, $i = 1, 2, \cdots, 30$. In this paper a procedure will be discussed for obtaining a Bayes estimate of the mean vector μ_i, the covariance matrix Σ_i, and the a priori probability q_i for the subclasses of interest and all of the subclasses in a scene. The procedure used to derive $\hat{\mu}_i$, $\hat{\Sigma}_i$, and $\hat{q}_i$ is to minimize m_L, which is the mean square error of the Bayes decision function of the single-class classifier.

*The material for this paper was developed under Contract NAS 9-12200 for the Earth Observations Division, Science and Applications Directorate, Lyndon B. Johnson Space Center, National Aeronautics and Space Administration, Houston, Texas.

III. BAYES ESTIMATION ON μ_i, Σ_i, AND q_i

By minimizing the mean square error of the Bayes decision function of the single-class classifier, mean vectors μ_i, covariance matrices Σ_i, and a priori probabilities q_i are derived as follows.

The Bayes decision function for the single-class classifier is

$$D(x) = \frac{\sum_{i=1}^{m} q_i p(x/i) - \frac{1}{2}p(x)}{p(x)} \qquad (3)$$

where $p(x)$ is the true mixture density function; m is the number of subclasses of wheat; and $\sum_{i=1}^{m} q_i p(x/i) - \frac{1}{2}p(x)$ is the discriminant function of the single-class classifier.

Let

$$\hat{D}(x) = \frac{\sum_{i=1}^{m} \hat{q}_i \hat{p}(x/i) - \frac{1}{2}\hat{p}(x)}{\hat{p}(x)} \qquad (4)$$

where $\hat{D}(x)$ is the estimated Bayes decision function.

Let

$$m_L = E_T\left\{[\hat{D}(x) - D(x)]^2\right\} \qquad (5)$$

where m_L is the mean square error of the Bayes decision functions with respect to total samples. By minimizing m_L, $\hat{\mu}_i$, $\hat{\Sigma}_i$, and $\hat{q}_i$ can be found.

Let

$$m_L' = m_L + K$$

$$K = \frac{1}{4} + \int \sum_{i=1}^{m} q_i p(x/i) dx = \text{constant} \qquad (6)$$

Minimizing m_L with respect to $\hat{\mu}_i$, $\hat{\Sigma}_i$, and $\hat{q}_i$ is equivalent to minimizing m_L' with respect to $\hat{\mu}_i$, $\hat{\Sigma}_i$, and $\hat{q}_i$. After performing some manipulation on m_L',

$$m_L' = E_T\left\{\left[\hat{D}(x) + \frac{1}{2}\right]^2\right\} - 2\sum_{i=1}^{m} q_i E_i\left[\hat{D}(x) - \frac{1}{2}\right] + K' \qquad (7)$$

where $K' = \int D(x)^2 p(x)dx$ is constant; $E_T[\ ]$ is the expected value of [ ] with respect to p(x), the mixture density function; $E_i[\ ]$ is the expected value of [ ] with respect to p(x/i), the probability density function of wheat subclass i; c is the number of subclasses of total samples; and m is the number of subclasses of wheat training samples.

To minimize m_L' with respect to $\hat{\mu}_i$, $\hat{\Sigma}_i$, and $\hat{q}_i$, the following partial derivatives are set to zero

$$\frac{\partial m_L'}{\partial \hat{\mu}_i} = 0 \quad ; \quad \frac{\partial m_L'}{\partial \hat{\Sigma}_i} = 0 \quad ; \quad \frac{\partial m_L'}{\partial \hat{q}_j} = 0 \qquad (8)$$

subject to the constraints

$$\sum_{i=1}^{c} q_i = 1 \text{ and } q_i > 0 \text{ for } \begin{cases} i = 1,2,\cdots,c \\ j = m+1,\cdots,c \end{cases} \qquad (9)$$

From equations (8) and (9), $\hat{\mu}_i$, $\hat{\Sigma}_i$, and $\hat{q}_i$ are derived as follows.

$$\hat{\mu}_i = \frac{E_T\left\{\left[\hat{D}(x) + \frac{1}{2}\right]\psi_i(x)x\right\} - \sum_{j=1}^{m} q_j E_j\left\{\psi_i(x)x\right\}}{E_T\left\{\left[\hat{D}(x) + \frac{1}{2}\right]\psi_i(x)\right\} - \sum_{j=1}^{m} q_j E_j\left\{\psi_i(x)\right\}} \qquad (10)$$

for $i = 1, 2, \cdots, c$.

$$\hat{\Sigma}_i = \frac{E_T\left\{\left[\hat{D}(x) + \frac{1}{2}\right]\psi_i(x)(x - \hat{\mu}_i)(x - \hat{\mu}_i)^T\right\} - \sum_{j=1}^{m} q_j E_j\left[\psi_i(x)(x - \hat{\mu}_i)(x - \hat{\mu}_i)^T\right]}{E_T\left\{\left[\hat{D}(x) + \frac{1}{2}\right]\psi_i(x)\right\} - \sum_{j=1}^{m} q_j E_j\left[\psi_i(x)\right]} \qquad (11)$$

for $i = 1, 2, \cdots, c$.

$$\hat{q}_i = \left[1 - \sum_{j=1}^{m}\hat{q}_j\right] \frac{2E_T\left\{\left[\hat{D}(x) + \frac{1}{2}\right]\phi_i(x)\right\} - 2\sum_{j=1}^{m} q_j E_j\left\{\phi_i(x)\right\}}{\sum_{i=m+1}^{c}\left[2E_T\left\{\left[\hat{D}(x) + \frac{1}{2}\right]\phi_i(x)\right\} - 2\sum_{j=1}^{m} q_j E_j\left\{\phi_i(x)\right\}\right]} \qquad (12)$$

for $i = m + 1, \cdots, c$ where

$$\psi_i(x) = \frac{\delta(i)\hat{q}_i\hat{p}(x) - \sum_{\ell=1}^{m}\hat{q}_\ell\hat{p}(x/\ell)\hat{q}_i}{\hat{p}(x)^2}\,\hat{p}(x/i)$$

$$\phi_i(x) = \hat{p}(i/x)\sum_{j=1}^{m}\hat{p}(j/x)$$

$$\delta(i) = \begin{cases} 1 & \text{for } i = 1, 2, \cdots, m \\ 0 & \text{for } i = m+1, \cdots, c \end{cases}$$

Equation (12) is used to estimate the a priori probabilities of nonwheat subclasses. For wheat subclasses, the historic a priori probabilities can be used.

To solve equations (10), (11), and (12), initial estimates are needed for $\hat{\mu}_L(0)$, $\hat{\Sigma}_i(0)$, and $\hat{q}_i(0)$. These can be obtained from statistics of wheat training samples and statistics from clustering the total samples. Applying these initial estimates and the iterating method[3] on equations (10), (11), and (12), $\hat{\mu}_i$, $\hat{\Sigma}_i$, and $\hat{q}_i$ can be found.

IV. SOME PRELIMINARY EXPERIMENTAL RESULTS

A simple experiment was conducted to investigate the convergence properties of the algorithm. Five hundred normally distributed samples were generated from each of two univariate distributions, as illustrated in figure 1, with means and variances of $\mu_1 = 10$, $\sigma_1^2 = 1.0$, $\mu_2 = 14$, and $\sigma_2^2 = 1.0$.

For this experiment, it was assumed that labeled samples were available for class 1 only. The samples for p(x) were obtained by forming a union of the samples generated for both classes with no labels attached to the samples. An attempt was made to estimate the mean of class 2 by using equation (10). The mean and variance of class 1 were set to $\hat{\mu}_1 = 10$, $\hat{\sigma}_1^2 = 1.0$, and the variance of class 2 was set to $\hat{\sigma}_2^2 = 1.0$. The prior probabilities were set to $\hat{q}_1 = 0.5$, $\hat{q}_2 = 0.5$. The initial value of the mean of class 2 was set to $\hat{\mu}_2(0) = 26$.

Figure 2 shows the value of $\hat{\mu}_2$ on successive iterations. Figure 3 shows the successive values of the mean square criteria [equation (7)]. In this example, m_L' converged to a minimum in two iterations. Figure 4 shows the probability of error for the successive mean values. The horizontal, dashed lines in these figures indicate the true or minimum values obtainable for each variable. Table 1 summarizes the final results obtained in iteration 2. In the example shown, the results indicate that for the mean, the convergence of the mean square criteria is rapid.

V. CONCLUSION

The single-class classifier which will classify data into the class of interest or the class "other" requires only training samples from the class of interest. This procedure minimizes the need for ground truth. This paper has presented a procedure for Bayes estimation on the parameters μ_i, Σ_i, and q_i of the single-class classifier.

In a simple, two-class example it was shown that the algorithm converges quite rapidly for the mean values only.

VI. REFERENCES

1. Minter, T. C.: The Single-Class Maximum Likelihood Classifier. Symp. on Machine Classification of Remotely Sensed Data, Laboratory for Applications of Remote Sensing, Purdue Univ. (West Lafayette, Ind.), June 3, 1975.

2. Lin, G. C.: The Single-Class Classifier. LEC/ASD Internal Memorandum 642-1694 (Houston), Oct. 1975.

3. Duda, R. O.; and Hart, P. E.: Pattern Classification and Scene Analysis. Wiley-Interscience, 1973, pp. 192-201.

Table 1. Results of Iteration 2

Estimated value of $\hat{\mu}_2$	True value of μ_2	Final value of m'_L	Value of m'_L at $\mu_2 = 14$ m'_L (minimum)	Probability of error for $\hat{\mu}_2 = 13.27$, percent	Minimum probability P_e (minimum), percent
13.27	14.0	0.522	0.515	2.98	2.27

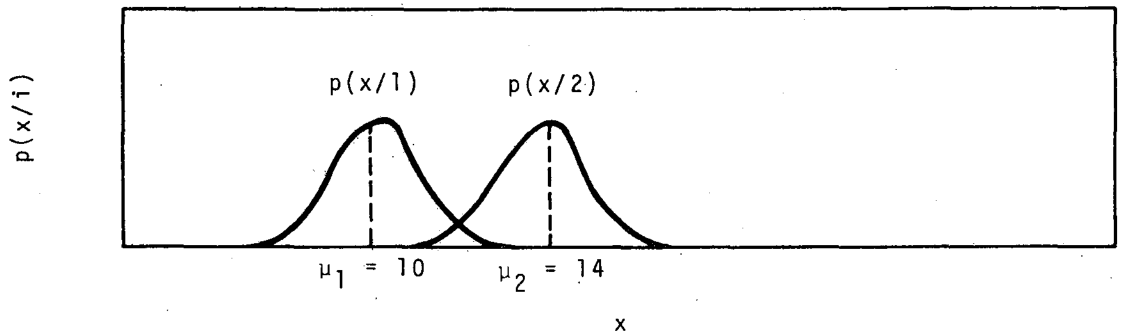

Figure 1. Two Univariate Distributions.

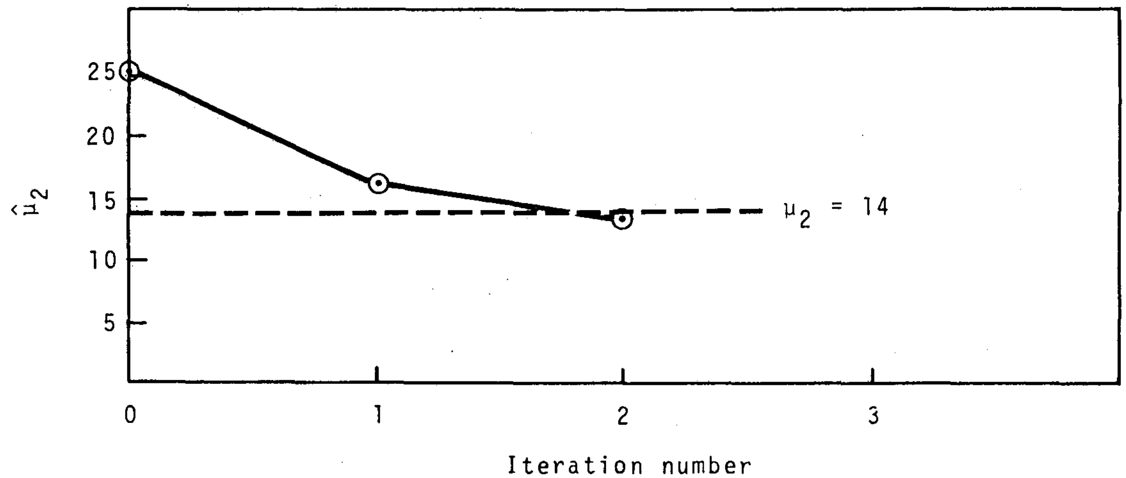

Figure 2. Mean of Class 2, μ_2.

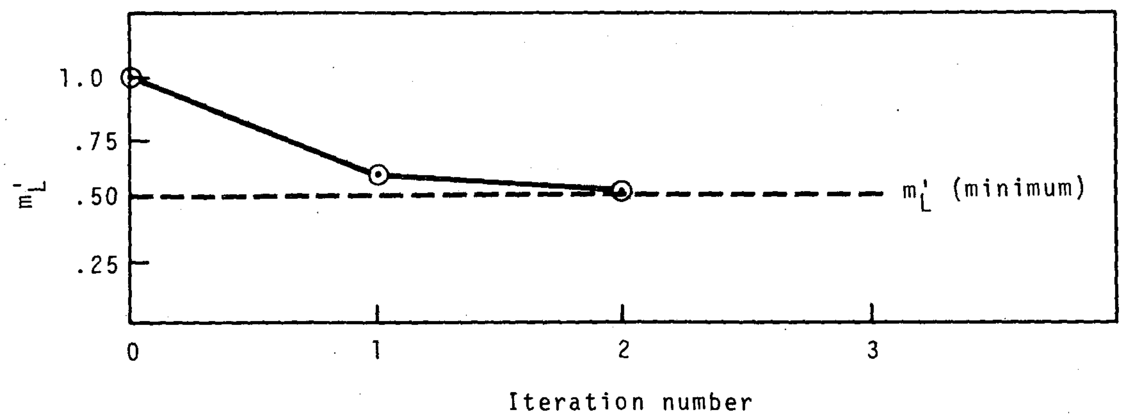

Figure 3. Value of Mean Square Criteria, m_L'.

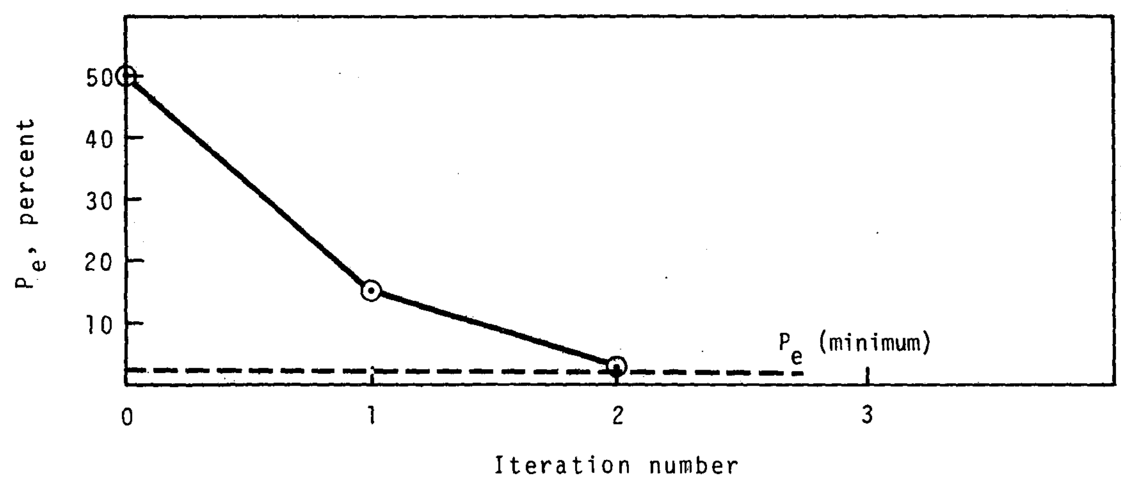

Figure 4. Probability of Error, P_e.